NBER WORKING PAPER SERIES

STEREOTYPE THREAT AND THE STUDENT-ATHLETE

Thomas S. Dee

Working Paper 14705
http://www.nber.org/papers/w14705

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2009

I would like to thank the Mellon Tri-College Forum for financial support through its seed grant program. I would also like to thank participants at the Fall 2008 NBER Higher Education Working Group meeting, the Tri-College Summer Seminar and the Mellon 23 Workshop "Evaluating Teaching and Learning at Liberal Arts Colleges" for useful comments. I would also like to thank Carolyn Abott, Andrew Fieldhouse, and Yimei Zhou for excellent research assistance. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Stereotype Threat and the Student-Athlete
Thomas S. Dee
NBER Working Paper No. 14705
February 2009
JEL No. C9,D0,I2

ABSTRACT

Achievement gaps may reflect the cognitive impairment thought to occur in evaluative settings (e.g., classrooms) where a stereotyped identity is salient (i.e., stereotype threat). This study presents an economic model of stereotype threat that reconciles prior evidence on how student effort and performance are influenced by this social-identity phenomenon. This study also presents empirical evidence from a laboratory experiment in which students at a selective college were randomly assigned to a treatment that primed their awareness of a stereotyped identity (i.e., student-athlete). This treatment reduced the test-score performance of athletes relative to non-athletes by 14 percent (effect size = -1.0).

Thomas S. Dee
Department of Economics
Swarthmore College
Swarthmore, PA 19081
and NBER
dee@swarthmore.edu

1 Introduction

The relevance of an individual's social identity for a variety of outcomes has long been a central concept in most social sciences. However, economists have only recently begun to explore explicitly the behavioral and welfare implications of social identities (Akerlof and Kranton 2000,2002,2005; Fang and Loury 2005; Bénabou and Tirole 2006; Benjamin, Choi, and Strickland 2007). One particularly prominent thread in the broader research on social identity has been the social-psychology literature on the phenomenon known as "stereotype threat." Stereotype threat refers to the perceived risk of confirming, through one's behavior or outcomes, negative stereotypes that are held about one's social identity. More specifically, its key conjecture is that the threat of being viewed through the lens of a negative stereotype can create an anxiety that disrupts cognitive performance and influences outcomes and behaviors. The seminal study on this topic (Steele and Aronson 1995) focused on the role of race in education-related settings and presented evidence that subtly priming experimental subjects to be aware of their racial identity did compromise their cognitive performance. Numerous studies have since replicated the effects of stereotype threat both with respect to social identities other than race (e.g., gender) and with respect to multiple outcomes (e.g., blood pressure, heart-rate variability, performance expecations, effort, risk aversion, and time discounting).

In this study, I present a simple economic model of stereotype threat. This model extends the conventional social-psychology explication of stereotype threat by explicitly modeling stereotype threat in a choice-based framework. In particular, this simple model reconciles seemingly contradictory results in the extant empirical literature by illustrating how the effects of stereotype threat on effort and performance depend on context-specific factors. This model also differs from recent economic models of social identity in a straightforward but conceptually

important detail. Previous economic models of identity have viewed individuals as choosing behaviors that correspond to the social norms for an identity (or set of identities) that they unambiguously view as their own. However, individuals experiencing stereotype threat do not necessarily feel that they personally do (or should) subscribe to stereotyped traits. Rather, it is the apprehension that *others* view them through the lens of a negative stereotype that is conjectured to create anxiety that compromises cogntive functioning and, eventually, identification with the stereotyped domain.

This study also presents the results of a stereotype-threat experiment that focuses on a particular social identity: that of a student-athlete at a selective post-secondary institution (i.e., Swarthmore College). The role of athletics at selective colleges and universities has recently received renewed critical scrutiny following the publication of two widely discussed books ("The Game of Life" by Shulman and Bowen [2001] and "Reclaiming the Game" by Bowen and Levin [2003]), which documented a growing "athletic-academic divide." Using the unique "College and Beyond" data set on students from 30 selective colleges and universities (including Swarthmore College), Shulman and Bowen (2001) presented evidence that student-athletes received a substantial admissions advantage but underperformed academically, even relative to their weaker academic preparation. The College Sports Project, a collaborative initiative of more than 70 Division III colleges and universities, has provided more current data on the comparative academic performance of student-athletes by tracking outcomes among the nearly 40,000 who entered the participating institutions during the 2005-06 academic year. The first year of data from this study indicate that there were only modest differences in the performance of female athletes and non-athletes but that male athletes (and, in particular, *recruited* male athletes) underperformed relative to non-athletes (i.e., 8 percentile points

lower in class rank).

Bowen and Levin (2003, page 165) argue that the academic underperformance of student-athletes at selective institutions is explained primarily by increasingly intense and specialized recruitment and admissions practices that unintentionally select for students with comparatively low levels of academic engagement (and not by the time demands or the unobserved socioeconomic traits of student-athletes). However, Shulman and Bowen (2001) also acknowledge that stereotype threat (i.e., the anxiety created by the threat of being viewed as an especially undeserving student in a highly selective academic setting) may also contribute to the academic underperformance of student-athletes. They also note that "unfortunately, we have no way of assessing how important stereotype threat is in the case of athletes" (Shulman and Bowen 2001, page 236). The experimental evidence presented here addresses this question directly by testing whether athletics-related stereotype threat compromises cognitive performance at one of the institutions in the "College and Beyond" data set.

2 Stereotype Threat and College Athletics

2.1 An Economic Model of Stereotype Threat

The model introduced in this section illustrates the conjectured mechanisms by which stereotype threat can influence both student effort and their level of skill formation. The basic theoretical framework introduced by Akerlof and Kranton (2002) extends their seminal economic analysis of identity (Akerlof and Kranton 2000) to issues related directly to schooling and provides a useful point of departure for this model. Akerlof and Kranton (2002) present models in which students explicity choose their social identity (e.g., leading crowd, nerd or burnout). In models of this type, utility is determined by the social status of

one's chosen identity and by how well an endowment of traits (e.g., appearance, intelligence) and a chosen level of effort allow one to approximate the ideal of that chosen social identity. Akerlof and Kranton (2002) discuss how these intriguing models provide new insights into several stylized facts and paradoxes from prior research on education policy. However, models of this type do not correspond exactly with how social psychologists conceptualize the interaction of social identity and stereotype threat. In particular, stereotype threat is not about conforming to the ideals of what an individual actor perceives as their salient social identity. Instead, the salient feature of stereotype threat is the apprehension and diminished cognitive performance that may be created by the suspicion about how one is viewed by *others*. In other words, the effects of stereotype threat do not depend on believing that one's identity corresponds to any particular stereotype. Rather, it is due to the anxiety from believing others may see you as belong to that stereotyped identity.

A simple extension of their model illustrates how stereotype threat may influence student effort and outcomes. In a baseline economic model of effort, an individual's utility reflects the return to skill, w, and a level of skill, $k(n, e)$, that is a function of ability, n, and effort, e:

$$wk(n, e) - \frac{1}{2}e^2 \tag{1}$$

The second term in this utility function reflects the disutility of expending effort. Also following Akerlof and Kranton (2002, page 1174), skill formation is assumed to increase linearly in one's endownment of ability (i.e., $k(n, e) = ne$). Akerlof and Kranton (2002) extend this simple model by introducing additional arguments that reflect the returns and costs associated with particular social identities that are available to choose. However, stereotype threat is not directly about the consequences of choosing an identity. Instead, the key feature

of stereotype threat is the cognitive disruption from situational threats due to concern about how one is viewed by *others* (e.g., being an athlete in classroom at a highly selective institution). This simple model can be adjusted to reflect this conceptualization by making the ability term, n, a decreasing function of situational threats, t, that create this anxiety (i.e., $n_t < 0$). Specifically, a simple model of stereotype threat would assume that an individual chooses a level of effort to maximize the following:

$$wn(t)e - \frac{1}{2}e^2 \tag{2}$$

In this extension of the baseline model, the effect of stereotype threat on the chosen level of effort, $e*$, can be shown to be unambiguously negative:

$$\frac{\partial e^*}{\partial t} = wn_t < 0 \tag{3}$$

The negative effect of stereotype threat on effort follows from the defining assumption that stereotype threat decreases the productivity of effort (i.e., $n_t < 0$). The intuition here is straightforward: stereotype threat is, in effect, a negative productivity shock that compromises the return to academic effort, thereby reducing the amount of effort chosen. The effect of stereotype threat on the equilibrium level of skill formation is illustrated by the following total derivative:

$$\frac{dk(n, e^*, t)}{dt} = e^* n_t + n\frac{\partial e^*}{\partial t} \tag{4}$$

The decision to expend less effort in the presence of stereotype threat ($\frac{\partial e^*}{\partial t} < 0$), combined with the reduction in the productivity of that effort ($n_t < 0$), implies that skill formation becomes unambiguously lower.

One notable shortcoming of this basic model is that it incorporates neither the disutility that students may experience directly from stereotype threat nor

how they respond to this disutility. Experiencing stereotype threat is likely to reduce individual utility *directly* by increasing anxiety and discomfort (i.e., not only through its effect on skill formation). Furthermore, the direct disutility created by stereotype threat could also have implications for the level of effort chosen. Individuals may attempt to reduce the direct disutility created by stereotype threat by increasing their effort in an attempt to deflect the stereotyped social identity. One way to think of this is as a "I'll show them!" response to the situational threat. A limited amount of empirical evidence suggests the possible relevance of this phenomenon. For example, Steele and Aronson (1995, page 804) found that a race-based prime increased the likelihood that black participants would answer survey questions in a manner that sought to avoid conforming to stereotypical images. Three gender-based studies suggest that stereotype threat could *increase* effort and motivation even by an amount sufficient to improve performance despite the reduced productivity of effort. Oswald and Harvey (2000) found that female undergraduates exposed to a hostile cartoon prior to taking a math test actually performed *better* when there was no attempt to reduce stereotype threat. Similarly, in a study of a visual-attention task, Jamieson and Harkins (2007) found that gender-related stereotype threat *increased* the motivation and performance of women when the experimental setting facilitated the opportunity to correct mistakes through increased effort (i.e., by allowing additional time). Third, a recent study by three economists (Fryer, Levitt, and List 2008) found that, despite reporting higher levels of stress, females performed at their best when primed to be aware of math-related gender stereotypes.

The model introduced here can be extended to reflect this particular dimension of stereotype threat by introducing a term that reflects the direct disutility associated with that threat, $S(t, e)$. This disutility is assumed here to be both

an increasing function of situational factors that create threat (i.e., t) and a decreasing function of effort. More specifically, this disutility is assumed here to take the following form - $S(t,e) = s(e)t$ - where the assumption that effort can reduce the disutility of feeling stereotyped implies that $s_e < 0$. The conjecture here is that increases in effort can reduce the disutility of experiencing stereotype threat by creating a more individual social identity that is viewed as belonging in the stereotyped domain.[1] In this extended model, individuals choose a level of effort to maximize:

$$wn(t)e - \frac{1}{2}e^2 - s(e)t \tag{5}$$

The effect of stereotype threat on the equilibrium choice of effort in this model (i.e., e^*) can be shown to take the following form:

$$\frac{\partial e^*}{\partial t} = \frac{wn_t - s_e}{1 + s_{ee}t} \gtreqless 0 \tag{6}$$

The denominator of this expression is unambiguously positive by the second-order condition. Therefore, whether stereotype threat actually increases effort (and, by implication, skill formation) can be seen to depend on the comparative magnitudes of wn_t and s_e. In other words, if stereotype threat leads to a relatively large reduction in the productivity of effort (i.e., $|wn_t|$ is large), then effort will decrease. However, if effort is relatively effective at reducing the disutility of experiencing stereotype threat (i.e., $|s_e|$ is large), then stereotype threat would *increase* the chosen level of effort.

These theoretical results illustrate the potentially complex and *context-specific* ways in which stereotype threat may influence effort and skill formation. The canonical theoretical prediction and empirical result in the stereotype-threat lit-

[1] Alternatively, this effect could be framed in terms of the expression for skills formation, $k(n,e,t)$, The results of doing so are similar to those discussed here.

erature is that the presence of this threat compromises performance through cognitive disruptions that effectively compromise the productivity of effort. However, some psychological studies have documented threat-induced *increases* in effort and, even, performance. The model presented here is consistent with this disparate findings. Specifically, this model indicates that, in social contexts where the direct disutility of experiencing stereotype threat is both large but readily reduced through effort, it is possible that stereotype threat will increase effort. A comparatively modest increase in effort would imply that skill formation is still lower in the presence of stereotype threat (i.e., the total derivative in equation [4] would still be negative) due to the reduced productivity of that effort. However, sufficiently large threat-induced increases in effort could even improve performance (Oswald and Harvey 2000, Jamieson and Harkins 2007, Fryer, Levitt, and List 2008). These findings could have particular relevance for stereotype-threat effects related to traits like athletic status relative to those related to race or gender. For example, it may be that one's social identity as an athlete (and the corresponding experience of stereotype threat) is more malleable in response to changes in effort than social identities and stereotype threat linked to race or gender. In situations like this (i.e., $|s_e|$ is large), threat-induced increases in effort are more likely to occur. The empirical evidence presented in this study is somewhat consistent with this (i.e., a threat-induced reduction in performance combined with some weakly suggestive evidence of an increase in effort).

2.2 Prior literature

Does stereotype threat contribute to the academic underperformance of student-athletes? Bowen and Levin (2003, page 164) note that there is a stigma attached to athletic participation at some selective institutions (especially in colleges and

in specific "high-profile" sports) and they suggest that the stereotype threat associated with this stigma could contribute to the academic underperformance of student-athletes. But they also argue that it is unlikely to explain the large differential fully, particularly in contexts where the athletic identity of a student may not be evident to a professor (e.g., larger universities, lower-profile sports and women's sports). However, this assertion of limited relevance turns on the assumption that stereotype threat is due exclusively or largely to concerns about how one is viewed by a professor. In fact, college students may care substantially more about their social identity as viewed by their *peers*, who are more likely to know their athletic status (at least at smaller, liberal-arts colleges). Because of these issues, the potential contribution of stereotype threat to the academic underperformance of athletes is a decidedly empirical question and one for which heterogeneity by institution traits (e.g., size and degree of athletic stigma) and student traits (e.g., sport played) are likely to play a role.

Most of the empirical evidence in support of the stereotype-threat phenomenon comes from laboratory experiments in which student-participants are randomly assigned to receive a treatment that "primes" their awareness of a stereotype prior to completing a test or some other task.[2] For example, in the seminal laboratory study by Steele and Aronson (1995), participating students were randomly assigned to be told that the test they were about to take was diagnostic of their ability (i.e., the stereotype-threat prime) or that the test was non-evaluative (i.e., the control condition). They found that black students in the "ability-diagnostic" condition performed significantly worse on tests than those in the control condition while the performance of white students was not significantly affected by how the test was framed. In another widely used variant of this study design, participants would first complete a brief questionnaire

[2] See Aronson and Steele (2005) for a discussion of this growing literature. Interestingly, there is also a small but growing body of *field* evidence that finds effects from interventions designed to buffer students from the effects of stereotype threat (e.g., Cohen et al. 2006).

that included questions designed to prime their awareness of a racial or gender identity (e.g., Steele and Aronson 1995, Shih, Pittinsky, and Ambady 1999, Benjamin, Choi, and Strickland 2007).

Only one prior study appears to have examined whether *athletics-related* stereotype threat influences cognitive performance. Yopyk and Prentice (2005) recruited participants from the football team, the ice-hockey team, and a-cappella singing groups at Princeton University (n = 67). Prior to taking a 10-question math test, these participants completed a questionnaire that included one of three possible manipulations: an extracurricular prime, a student prime, and no prime. Athletes in the extracurricular prime condition were asked to write about their last athletic competition while the singers were asked to write about their last performanc e. In the student-prime condition, students were asked to write about their last academic success. In the no-prime condition, students were asked to write out explicit directions for getting from their dorm room to the main library. Yopyk and Prentice (2005) found that athletics assigned to the extracurricular prime were substantially less accurate on the math test while there was not a similar effect for the a cappella singers.[3]

These results suggest that stereotype threat contributes to the academic underperformance of student-athletes even at a larger institutions where this social-identity phenomenon might have been thought less relevant. However, there are several reasons to be circumspect about the results of this study. First, the fact that participation in athletics was a condition for recruitment in the study could compromise the external validity of these findings. Students who were explicitly recruited for the study on the basis of their athletic status may have been exceptionally susceptible to the threat prime. Furthermore, the

[3] The size of this statistically significant effect is quite large. The regression-adjusted percent correct among athletes was 86 percent in the student-prime condition but only 67 percent in the extracurricular-condition. Interestingly, the extracurricular prime led both singers and athletes to answer fewer questions.

priming mechanism (i.e., writing detailed comments on athletic participation) is also unusually strong and does not parallel how a social identity as an athlete is likely to be manifested in real-world settings. Another issue with respect to external validity is that this study only only included males from two sports that are considered high-profile.

A particularly relevant concern with this study concerns internal validity. Yopyk and Prentice (2005) used simple randomization to assign each of the 67 participants to one of the three experimental conditions. However, a "failure of randomization" occurred such that participants in the student-prime condition had significantly SAT scores than other participants. Yopyk and Prentice (2005) suggest that their analysis corrects for this problem by conditioning on SAT scores. However, conditioning on observables is unlikely to provide a satisfactory solution for a randomization "failure." This is because the high-SAT students who were predominantly in the student-prime condition were likely to have any number of unobserved traits that influenced their test performance. Furthermore, the likely direction of this bias is one that would confound this study's main finding. Those assigned to the student prime may have outperformed those in the extracurricular prime simply because they had unobserved traits that predisposed them to do so.

3 A Stereotype-Threat Experiment

The laboratory experiment described here tests whether stereotype threat appears to contribute to the academic underperformance of college student-athletes by implementing a conventional stereotype-threat experiment but with the distinction that an athletic identity, rather than identities related to race or gender, is primed. This study was also designed to improve upon the limited prior evidence (i.e., Yopyk and Prentice 2005) in two potentially critical ways. First,

in order to increase the external validity that laboratory results might have for real-world educational settings, the recruited partcipants were not aware of the athletic focus of the study. Furthermore, the "priming" mechanism used in this study and described below was both more subtle and more likely to reflect a way in which athletic identity is actually made salient in classroom settings. Second, this study addresses "internal validity" concerns about the potential failure of simple randomization to create balance across the treatment and control states by utilizing a simple "block randomization" procedure (i.e., pairing participants on relevant baseline traits and randomizing within pairs) that is described below.

As with any laboratory experiment, the external validity of these results for other populations and real-world settings is an important issue. For example, there are at least two reasons to suspect that a study of this type may have unique power when based on students at Swarthmore College. First, the small size of the College suggests that student-athletes are particularly likely to view their athletic status as well known. Second, a fairly long history of animus in the College community with respect to the relationship between athletics and the core academic mission of the College also suggests that a definite athletic stigma exists in the community. However, the broader criticisms of relevance that are often raised about laboratory experiments may have unusually low applicability in this particular context.[4] This is partly because college students are the exact study population of interest in this application. In other words, this is not a case of using only college students to draw inferences about the nature of altruism in the broader society or the strategic behavior of firms in imperfectly competitive markets. Furthermore, in this application, the unusual level of scrutiny that characterizes the laboratory environment as well as the focus of the lab activity (i.e., cognitive performance) correspond in an uniquely direct manner with the

[4] See Levitt and List (2007a,b) for analyses and discussion of these concerns.

broader research questions of interest (i.e., the performance of college students in the similarly evaluative context of academic classrooms).

3.1 Recruitment

At the beginning of the Spring 2008 semester, all students at Swarthmore College received emails inviting them to participate in a 1-hour laboratory experiment whose goal was "to examine the determinants of cognitive functioning." Students were also told that they would receive $15 for their hour of participation. In order to promote statistical power and generalizability, additional emails and a study-recruitment letter were sent to student-athletes. However, none of the recruitment materials indicated the athletic focus of the study. Furthermore, students could not directly infer from the emails or the mailer this selective recruiting strategy. Current student-athletes were identified through the rosters that were publicly available on the College athletics web site.[5] All recruited students were directed to a secure web page where they could register for the study by completing a questionnaire and indicating their scheduling availability. The brief baseline questionnaire included questions about the participant's age, graduating class, academic major, sex, and math and verbal SAT scores.[6]

Ninety-one students completed this registration and were randomized according to the procedure described below. These participating students could then select into one of five scheduled laboratory sessions that were held in the 5th and 6th weeks of the semester. Seven students who had registered for the

[5] Students-athletes who no longer played a sport could not be identified from these data. However, current and past athletic participation was later elicited either in the beginning of the experiment (i.e., the treatment students) or as part of the exit questionnaire (i.e., the control students).

[6] Conversions to SAT scores were made for two of the participating students for whom only ACT scores were available. A small number of students could not recall their SAT scores. With permission, SAT scores were obtained for those students from the Registrar's office. In order to comply with the informed-consent procedures, students who were younger than 18 were not included in the study.

study did not utlimately attend a session leaving a final sample of 84.[7]. Roughly a quarter of students at the College were athletes. However, 44% of the study participants (i.e., 37 out of 84) were athletes, which reflected the success of the differential recruitment strategy. The presence of non-athletes in the study makes it possible to separate the effects of the stereotype-threat intervention unique to athletes from any unintended, general effect it might have. In any experiment such as this one, an important question involves the other ways in which the participants do or do not resemble the larger population from which they were drawn. Table 1 provides just such a comparison. The study population actually resembled the overall student body quite closely with respect to gender and SAT scores. However, study participants were both more likely to be freshman instead of seniors and less likely to be black or Hispanic.

3.2 Randomization

In order to increase the likelihood that the unobserved participant traits were unrelated to treatment status, the randomization procedure used in this study exploited the baseline traits available from the initial questionnaire and the publicly available athletics rosters. Participants were matched to other participants with respect to traits thought to be relevant to the study (e.g., athletic status, math SAT scores). Randomization then occurred within these matched pairs. Students who were identified as current athletes from the roster data were first sorted into cells based on the gender-specific sport that they played (e.g., women's soccer, men's basketball). The choice to group the athletes by genders-specific sports reflected the concern that the existence and degree of stereotype threat could be sport-specific as well as the fact that the sport played may proxy for unobservables related to achievement. For the same reasons, multi-sport

[7] The characteristics of the attriters and the balance of the participants by treatment status are discussed below.

athletes were grouped with other students playing similar sport combinations. Within each of these sport-gender cells, each participant was matched to another participiant with similar math SAT scores. In cases where there were an odd number of participants within cells, participants were matched to someone of the same gender and similar math SAT score but a different sport. Any remaining participant was assigned a treatment status through simple randomization. The participants who did not appear on the current athletics rosters were assigned a treatment status by a similar procedure. Specifically, they were first sorted by gender and graduating class. Then participants were paired within each class-gender cell to a another participant with a similar math SAT score. In cases where there were odd remainders in a gender-by-cohort cell, they were matched with a similar residual from a neighboring gender-specific cohort if available or randomized as a singleton otherwise.

The fundamental goal of randomization is, of course, to increase the likelihood that any effects associated with the experimental treatment reflect the true impact of that treatment rather than participant traits that may merely correlate with the treatment. However, the intent of randomization can be compromised both by non-random attrition from the study as well as by chance. Given the design features of this study, neither concern is likely to be relevant in this context. More specifically, because participants were unaware of their treatment status, it should be unrelated to attrition. Furthermore, the baseline matching procedure described above should dramatically reduce the chance of a randomization "failure." Nonetheless, auxiliary regressions provide a straightforward framework for assessing the ex post relevance of these concerns. Table 2 reports the key results from linear probability models where attrition from the study is the dependent variable.[8] Of the 7 students who attrited 5 had been assigned to the treatment condition. However, the results in Table 2 indicate

[8] Probit and logit models return results similar to those reported here.

that treatment status was not significantly related to attrition. These results do indicate that minority students were significantly more likely to attrit and that reading SAT scores were also positively associated with attrition. Table 3 reports the results of linear probability models where treatment status is the dependent variable. The results are uniformly consistent with a "successful" randomization in that treatment status is unrelated to athletic status, gender, graduating class, race-ethnicity, and SAT scores.[9] In fact, in each of these specifications, these regressors are jointly insignificant as well. A test statistic based on the null hypothesis that $R^2 = 0$ has a probability value of 0.52 or higher.

3.3 Experimental Procedures

Based on their availability, the study participants attended one of five 1-hour lab sessions that occurred in the 5th and 6th weeks of the spring 2008 semester. These 1-hour lab sessions were conducted in normal College classrooms by an administrator who was blind to the treatment status of the participants. Before beginning the experiment, the participants first reviewed and signed an informed consent form. The administrator then distributed individual-specific folders of materials to each participant. The experiment then consisted of guiding the students through the sequenced completion of five distinct sets of experimental materials (i.e., questionnaires, a test, etc.) in the folders. The experiment began with a 1-page questionnaire. For students in both the treatment and the control states, the questionnaire elicited information on the student's graduating class, whether they lived in College housing and whether they had a roommate.

For the students in the treatment condition (both athletes and non-athletes), the questionnaire then asked "Are you (or have you been) a member of a National Collegiate Athletics Association (NCAA) sports team at the College?". They were then asked to identify the sport(s) they played and to respond to three

[9] Again, probit and logit models return results similar to those reported here.

questions about the frequency with which (on a scale of 1 to 7) they experienced scheduling conflicts between athletics and, respectively, course/seminar meetings, laboratory sessions, and other academic lectures (e.g., evening lectures by outside speakers). For students in the control condition, the questionnaire continued instead with similarly structured questions related to the dinings services on campus. The basic structure of these treatment and control questionnaires parallels those used in the stereotype-salience study by Shih, Pittinsky, and Ambady (1999) and a recent study by Benjamin, Choi, and Strickland (2007).

Following the completion of this brief questionnaire, the participants were instructed that they would have 30 minutes to complete a 39-question test. The administrator explained to the participants that they might not be able to finish the test in the allotted time but that they should try to answer correctly as many questions as possible. The test consisted of 30 quantitative questions and 9 verbal questions from a Graduate Record Examination (GRE). As in prior studies of stereotype threat, this study reports the effect of random assignment to the stereotype-threat prime on participants' test accuracy (i.e., the percent correct of answered questions) and on the number of questions answered. The stereotype-threat prime could influence the test accuracy of participants through its effects on both cognitive functioning and test effort (i.e., respectively, the n and e terms in the theoretical model). The number of questions answered is commonly used as a less ambiguous proxy for participant effort (e.g., Fryer, Levitt, and List 2008). The extent to which the number of answered questions is actually a reliable proxy for participant effort is unknown. However, the instructions deliberately encouraged participants to care both about the accuracy of their answers and the number of completed questions. And the data on completed questions suggest that participants did apply significant effort on this margin. The 30-question quantitative section of the test was designed by the

Educational Testing Service (ETS) to be taken though not necessarily completed in a 30-minute GRE session. So, the addition of 9 additional GRE questions to a 30-minute test period constituted an unusually stringent hurdle in terms of answering questions. Nonetheless, slightly more than a third of participants answered all 39 questions; 80 percent answered 30 questions or more and no participant answerd fewer than 22 questions.

At the conclusion of the 30 minutes allotted for the test, the students were then directed to a word completion exercise designed to test the cognitive activation of the stereotype. Specifically, this exercise consisted of 30 word fragments, 12 of which were designed with the possibility that they could be completed as sports-themed words (e.g., "GO _ _" which could be completed as "GOAL"). This list also contained word fragments that could be completed in a way that suggested self doubt (e.g., "DU _ _" as "DUMB") and 11 filler words. After this 10-minute exercise, the students were directed to a short questionnaire that consisted of the 7 questions that constitute the academic sub-scale of the self-regard survey (Fleming and Courtney 1984). The experiment then concluded with a short exit questionnaire where participants could indicate the extent to which they enjoyed the study and what they thought the study's purpose was. All students were also asked to identify their race-ethnicity. For students assigned to the control condition, the exit questionnaire then contained the questions about athletics that had been in the opening questionnaire for students in the treatment condition. For students in the treatment condition, the questionnaire instead continued with the questions related to dining services.

4 Results

On average, the participating students answered 35 of the 39 available questions with 27 questions answered correctly. The primary measure of test performance,

the percent of answered questions that were correct, averaged 78.4 percent with a minimum value of 43 percent, a maximum value of 97 percent and a standard deviation of 0.11. Figure 1 presents graphical evidence on the effects of the intervention by showing kernel-density estimates of the test-performance distributions by treatment status and athletic status. The top panel of Figure 1 indicates that, for the non-athletes, the distributions of test-score performance are remarkably similar by treatment status.[10] In contrast, the bottom panel of Figure 1 indicates that, for the athletes participating in the study, assignment to the threat condition led to a quite large leftward shift in the test-performance distribution, an effect consistent with the hypothesis of stereotype threat.

The regression specification used to estimate the effect of the intervention on test performance (i.e., y_i) takes the following form:

$$y_i = \alpha + \beta(T_i A_i) + \gamma T_i + \delta A_i + \boldsymbol{\theta}\mathbf{X}_i + \varepsilon_i \tag{7}$$

where T_i and A_i are binary indicators that identify, respectively, whether student i was assigned to the treatment and was an athlete and ε_i is a mean-zero random error term. The coefficient of interest, β, reflects the unique effect the stereotype-threat intervention had on athletes. The matrix, $\mathbf{X}_i$, contains various other determinants of test performance, including fixed effects for gender and race, math and verbal SAT scores, and fixed effects for the student's graduating class and laboratory session they attended. Given the random assignment, none of these control variables should have a substantive influence on the estimated value of β. However, these controls can improve the precision of this point estimate. The fixed effects for laboratory sessions provides a control for unintended determinants that may have been unique to each session (e.g., administrator

[10] These comparative distributions may appear to suggest that the treatment modestly increased test-score performance (e.g., a possible stereotype "lift"). However, this interpretation reflects just two or three outlying observations. Regression-adjusted comparisons indicate that the treatment did not have a statistically significant effect on the non-athletes.

behavior, classroom setting, peer traits, etc.).

The key results from estimating this model are reported in Table 4. The full-sample results indicate that the unique effect of the treatment on athletes (i.e., the estimate of β) on athletes was uniformly negative and implied reductions in test performance ranging from 7.3 to 9.5 percentage points. In specifications that include few or no controls (i.e., columns 1 and 2), this estimate was somewhat smaller and statistically significant or weakly significant. However, in estimates that include additional controls (in particular, SAT scores), this estimate becomes larger and more precise. In contrast, the estimated main effect of the treatment (i.e., the estimate of γ) was positive but smaller and statistically insignificant in all specifications. The results in Table 4 also indicate that SAT scores strongly predicted test performance and that female participants tended to perform somewhat worse.

The theoretical model introduced in this study suggested that contextual factors can mediate the influence of stereotype threat. One potentially important aspect of this possible heterogeneity involves whether the effects of stereotype threat vary with the sport played. Participation in particular sports (e.g., lacrosse) may involve a unique academic stigma. Interestingly, the data collected by the College Sports Project suggest that the comparative academic performance of student-athletes varies considerably by sport (e.g., male and female cross-country athletes *outperform* non-athletes academically). The number of athletes in this study (n = 37) does not allow for a statistically meaningful breakdown by sports. However, one indirect way to address this issue to re-evaluate the results from the first panel of Table 4 on the sub-sample of respondents that excludes athletes who participated in track and field or cross country (i.e., sports thought to have no academic stigma or, possibly, a positive one).[11] The results in the right panel of Table 4 are based on the resulting

[11] The observed traits of students in this subsample are unrelated to treatment status as in

sample of 75 respondents. Consistent with expectations, the estimated β becomes uniformly larger ranging from 0.109 to 0.118 and, across all specifications, statistically distinguishable from zero at the 5-percent level.

The results presented in Table 5 explore other types of response heterogeneity using specification (10) from Table 4 as a baseline. More specifically, Table 5 reports the estimated value of β based on samples defined by gender, race, math SAT scores and graduating class. While the sample sizes do not allow for much precision, at least two results are noteworthy. The treatment effect unique to athletes appears to be concentrated both among males and among respondents with math SAT scores below the sample median (i.e., 720).

Overall, these results suggest that stereotype threat may make a substantive contribution to the academic underperformance of student-athletes documented by Shulman and Bowen (2001). The stereotype-threat treatment studied here reduced test performance by 11 percentage points (i.e., specification (10) in Table 4), which is equivalent to a 14 percent reduction in the mean performance score or, alternatively, an effect size of roughly -1.0. Treatment effects of this size are not uncommon in laboratory tests related to stereotype threat (e.g., Steele and Aronson 1995, Shih, Pittinsky, and Ambady 1999, Yopyk and Prentice 2005). Effects of this magnitude as well as their relative concentration among males and higher-profile sports are also roughly consistent with the patterns of academic underperformance reported by Shulman and Bowen (2001) and the College Sports Project. For example, Shulman and Bowen (2001, Table 3.1) find that among males from the most recent "College and Beyond" cohort who attended cohort liberal arts colleges, the regression-adjusted effect associated with being an athlete in a high-profile sport is a 8.8 percentile point reduction in class rank. The corresponding effect for females was 6.1 percentile points.

The theoretical model of stereotype threat introduced in this study suggested

Table 3.

that the effects of stereotype threat on student effort and performance could be highly context-specific. For example, students may sometimes respond to stereotype threat with increased effort that may or may not be able to overcome threat-induced reductions in the productivity of effort that compromise task performance. The evidence from this experiment is weakly suggestive of this phenomenon. More specifically, given the ambiguous instructions to answer correctly as many questions as possible, the number of test questions answered can be viewed as a proxy for effort. As the results in Table 6 indicate, the stereotype-threat treatment *increased* the number of test questions answered by athletes by roughly 5 percent but *reduced* the number of correctly answered questions by 8 percent. However, these estimated effects fall short of statistical significance.[12]

The ancillary measures collected after the completion of the test were meant to capture the effects of the stereotype-threat intervention on awareness of athletics-related stereotypes as well as measures of self-doubt and academic self-regard. However, analyses of these measures based on specifications like that in equation (7) suggest that there were no statistically significant effects (Table 6). It should be noted that the test administered in this study was quite long relative to similar studies in terms of the number of questions and the corresponding length of the testing period (i.e., 30 minutes instead of 5). The motivation for this design was to avoid ceiling effects with respect to test performance. This design may have contributed to the absence of effects on the post-test measures if it left subjects too fatigued or stressed to give valid responses. One empirical result consistent with this interpretation is that virtually none of the observed student traits (e.g., race, gender, SAT scores) were significant predictors of the data collected in the last 20 minutes of the experiment

[12] The results are similarly imprecise in count-data and semi-log specifications as well as in models based on sub-samples defined by participant traits (e.g., males and below-median math SAT scores).

(i.e., academic self-regard, self-doubt or sports-themed word completions).

5 Conclusions

The prominent role that social identity may play in influencing a broad array of economic and education-related outcomes is receiving an increasing amount of attention. This study makes two broad contributions to this literature. One is to adapt an economic model of social identity to reflect more accurately one of the most prominent conjectures in the literature on social identity, the phenomenon of stereotype threat. This model illustrates how prior, seemingly contradictory, empirical evidence on how stereotype threat influences effort and task performance can be reconciled in a model that captures how individuals may respond to the direct disutility created by experiencing stereotype threat. The second goal of this study is to examine through a laboratory experiment whether stereotype threat contributes to a large and controversial achievement gap observed at many selective colleges and universities: the academic underperformance of student-athletes. The results of this study are consistent with the hypothesis that the academic stigma associated with being a student-athlete at a highly selective college or university makes a substantial contribution to their academic underperformance.

In the social sciences, the relevance for real-world settings of laboratory studies such as this is always suspect (Levitt and List 2007). This study is no exception. However, it is also the case that the laboratory experiment discussed here is likely to have an unusually degree of external validity. The study participants (i.e., college students), the key outcome measure (i.e., test performance) and the unusual evaluative scrutiny of a laboratory session all strongly parallel the field setting of interest (i.e., the academic environment at a selective institution). Nonetheless, a compelling next step for research in this area would involve

well-designed trials that assess both the apparent relevance of stereotype threat in a field setting and strategies to amerliorate this social-identity phenomenon (e.g., Cohen et al. 2006).

References

[1] Akerlof, George and Rachel Kranton. "Economics and Identity," *Quarterly Journal of Economics* 115(3), August 2000, pages 715-753.

[2] Akerlof, George and Rachel Kranton. "Identity and Schooling: Some Lessons for the Economics of Education," *Journal of Economic Literature* 40, December 2002, pages 1167-1201.

[3] Akerlof, George and Rachel Kranton. "Identity and the Economics of Organizations," *Journal of Economic Perspectives* 19(1), Winter 2005, pages 9-32.

[4] Aronson, Joshua and Claude M. Steele. "Stereotypes and the Fragility of Academic Competence, Motivation, and Self-Concept," in *Handbook of Competence and Motivation*, Andrew J. Elliot and Carol S. Dweck, editors, New York: Guilford Press, 2005.

[5] Bénabou, Roland and Jean Tirole "Identity, Dignity and Taboos: Beliefs as Assets," working paper, December 2006

[6] Benjamin, Daniel J., James J. Choi, and A. Joshua Strickland. "Social Identity and Preferences," NBER Working Paper No. 13309, August 2007.

[7] Bowen, William G. and Sarah A. Levin. *Reclaiming the Game: College Sports and Educational Values*. Princeton University Press, 2003.

[8] Cohen, Geoffrey, Julio Garcia, Nancy Apfel, and Alison Master. 2006. "Reducing the racial achievement gap: A social-psychological intervention," *Science* vol. (313), September 1: 1307-1310.

[9] Fang, Hanming and Glenn C. Loury "'Dysfunctional Identities' Can Be Rational," working paper, January 9, 2005.

[10] Fleming, James S. and Barbara E. Courtney. "The Dimensionality of Self-Esteem II:. Hierarchical Facet Model for Revised Measurement Scales," *Journal of Personality and Social Psychology* 46(2), 1984, pages 404-421.

[11] Fryer, Roland G., Levitt, Steven D. and John A. List. "Exploring the Impact of Financial Incentives on Stereotype Threat: Evidence from a Pilot Study," *American Economic Review*, May 2008.

[12] Jamieson, J. P., & Harkins, S. G. (2007). Mere effort and stereotype threat performance effects. *Journal of Personality and Social Psychology*, 93, 544-564.

[13] Levitt, Steven D. and John A. List. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21(2), Spring 2007, pages 153-174.

[14] Oswald, D. L., & Harvey, R. D. (2000). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology* 19(4): 338-356.

[15] Shih, Margaret, Todd L. Pittinsky, and Nalini Ambady. "Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance," *Psychological Science* 10(1), January 1999, pages 80-83.

[16] Shulman, James L. and William G. Bowen. *The Game of Life: College Sports and Educational Values*. Princeton University Press, 2001.

[17] Steele, Claude M. and Joshua Aronson. "Stereotype Threat and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology* 69(5), 1995, pages 797-811.

[18] Yopick, Darren J.A. and Deborah A. Prentice. "Am I an Athlete or a Student? Identity Salience and Stereotype Threat in Student-Athletes," *Basic and Applied Social Psychology* 27(4), 2005, pages 329-336.

Table 1 - Average Traits of College Students and Study Participants

		Study Participants			
Variable	College	Total	Treatment	Control	p-value
Athlete	24.4%	44.0%	46.3%	41.9%	0.68
Female	52.0%	52.7%	53.7%	53.5%	0.99
White Non-Hispanic	43.5%	61.5%	73.2%	60.5%	0.22
Asian	17.3%	22.0%	19.5%	27.9%	0.37
SAT (Math)	736	724	729	721	0.51
SAT (Reading)	716	718	720	714	0.68
Class of 2008	27.2%	18.7%	19.5%	18.6%	0.92
Class of 2009	24.7%	24.2%	22.0%	23.3%	0.89
Class of 2010	22.5%	22.0%	17.1%	25.6%	0.35
Class of 2011	24.5%	35.2%	41.5%	32.6%	0.40
Enrollment	1,491	84	41	43	

Source: The Fact Book, Institutional Research Office, Swarthmore College (http://www.swarthmore.edu/factbook.xml). College-specific athletic participation based on gender-specific 2005-06 data weighted by male and female enrollment. SAT scores are based on the matriculants from 2005, 2006, and 2007. The p-value refers to a test of the hypothesis that the prevalence of the observed trait is the same across the treatment and control groups.

Table 2 - Auxiliary Regressions, OLS Estimates of the Determinants of Study Attrition

Independent Variable	(1)	(2)	(3)	(4)
Treatment	0.064	0.054	0.053	0.053
	(0.056)	(0.043)	(0.043)	(0.043)
Athlete	-0.081	-0.064	-0.021	-0.030
	(0.057)	(0.045)	(0.050)	(0.051)
Female	-	0.012	0.007	0.013
		(0.043)	(0.043)	(0.045)
White Non-Hispanic	-	-0.459‡	-0.504‡	-0.506‡
		(0.060)	(0.064)	(0.065)
Asian	-	-0.470‡	-0.495‡	-0.479‡
		(0.071)	(0.076)	(0.078)
SAT (Math)	-	-	0.0001	0.000003
			(0.0005)	(0.0005)
SAT (Reading)	-	-	0.0008†	0.0007*
			(0.0004)	(0.0004)
Class of 2009	-	-	-	0.030
				(0.067)
Class of 2010	-	-	-	-0.052
				(0.071)
Class of 2011	-	-	-	-0.054
				0.063
R^2	0.0365	0.4464	0.4747	0.4916

Notes: The sample size is 91 of whom 7 were attriters. Five of the attriters had been assigned to the treatment.

*Statistically significant at the 10-percent level

†Statistically significant at the 5-percent level

‡Statistically significant at the 1-percent level

Table 3 - Auxiliary Regressions, OLS Estimates of the Determinants of Treatment Status

Independent Variable	(1)	(2)	(3)	(4)	(5)
Athlete	0.045	0.028	0.063	0.078	0.016
	(0.111)	(0.117)	(0.130)	(0.133)	(0.139)
Female	-	-0.001	0.003	0.032	0.085
		(0.113)	(0.113)	(0.119)	(0.123)
White Non-Hispanic	-	0.164	0.120	0.108	0.104
		(0.194)	(0.213)	(0.220)	(0.216)
Asian	-	0.037	-0.035	-0.074	-0.078
		(0.221)	(0.238)	(0.244)	(0.241)
SAT (Math)	-	-	0.001	0.001	0.001
			(0.001)	(0.001)	(0.001)
SAT (Reading)	-	-	-0.00001	0.0003	0.001
			(0.001)	(0.001)	(0.001)
Class of 2009	-	-	-	-0.057	-0.120
				(0.179)	(0.180)
Class of 2010	-	-	-	-0.074	-0.192
				(0.190)	(0.202)
Class of 2011	-	-	-	0.106	0.035
				(0.165)	(0.172)
Session Fixed Effects	No	No	No	No	Yes
R^2	0.0020	0.0190	0.0295	0.0509	0.1477
p-value (H_0: R^2=0)	0.6836	0.8204	0.8829	0.9085	0.5244

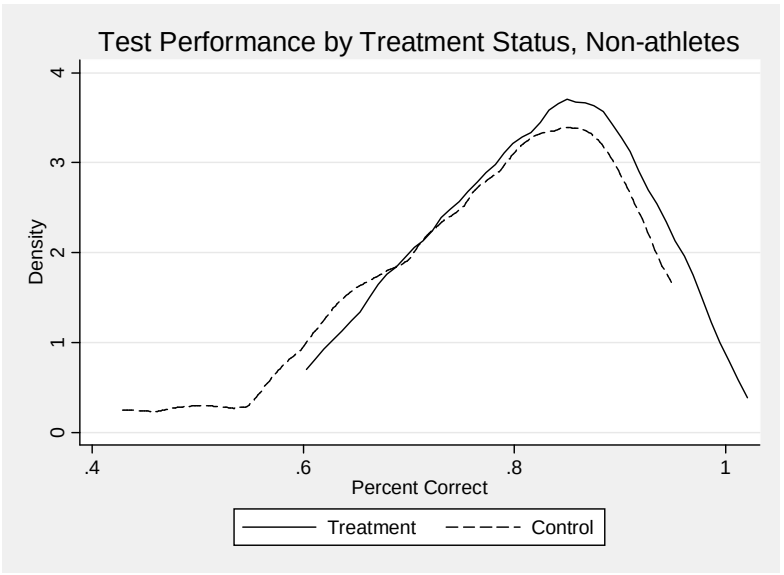
Notes: The sample size is 84 of whom 41 had been assigned to the treatment. The session fixed effects control for which of the 5 laboratory sessions the respondent attended.

*Statistically significant at the 10-percent level

†Statistically significant at the 5-percent level

‡Statistically significant at the 1-percent level

Figure 1 - Kernel Density Graphs of Test Performance
by Treatment and Athletic Status



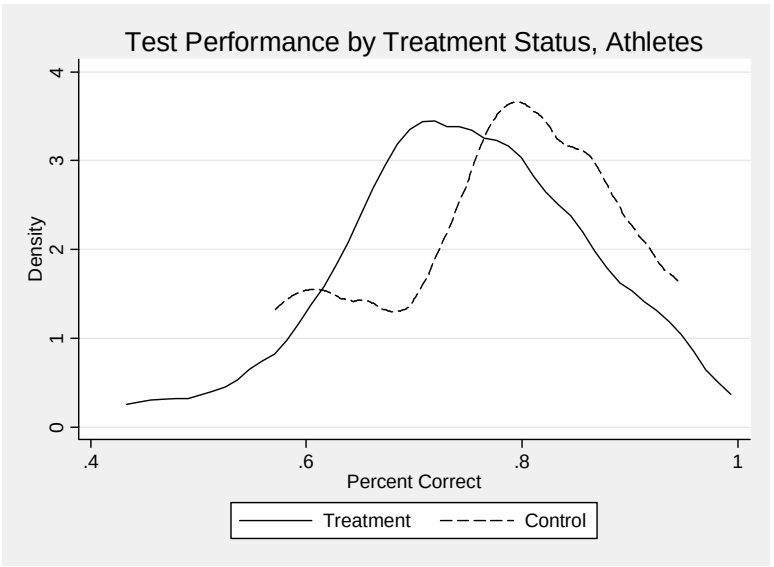


Table 4 - OLS Estimates of the Determinants of Test Performance

Independent	Full Sample (n=84)					Excluding Non-Stereotyped Sports (n=75)				
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment x Athlete	-0.073	-0.081*	-0.091†	-0.087†	-0.095†	-0.109†	-0.118†	-0.113‡	-0.112†	-0.110†
	(0.049)	(0.048)	(0.041)	(0.041)	(0.042)	(0.053)	(0.052)	(0.043)	(0.043)	(0.044)
Treatment	0.036	0.041	0.036	0.036	0.037	0.037	0.041	0.040	0.043	0.042
	(0.032)	(0.032)	(0.027)	(0.028)	(0.029)	(0.033)	(0.032)	(0.027)	(0.027)	(0.029)
Athlete	0.002	0.018	0.077†	0.067†	0.064*	0.020	0.036	0.097‡	0.089‡	0.086†
	(0.034)	(0.035)	(0.031)	(0.032)	(0.033)	(0.037)	(0.037)	(0.032)	(0.033)	(0.035)
Female	-	-0.049†	-0.048†	-0.044†	-0.044*	-	-0.041	-0.042†	-0.041*	-0.038
		(0.024)	(0.020)	(0.021)	(0.023)		(0.025)	(0.021)	(0.021)	(0.024)
White Non-Hispanic	-	0.042	-0.037	-0.047	-0.051	-	0.078	-0.023	-0.022	-0.030
		(0.042)	(0.038)	(0.039)	(0.040)		(0.052)	(0.046)	(0.047)	(0.049)
Asian	-	0.080*	-0.003	-0.004	-0.011	-	0.121†	0.027	0.037	0.026
		(0.047)	(0.043)	(0.044)	(0.044)		(0.055)	(0.050)	(0.051)	(0.053)
SAT (Math)	-	-	0.001‡	0.001‡	0.001‡	-	-	0.001‡	0.001‡	0.001‡
			(0.0002)	(0.0002)	(0.0002)			(0.0002)	(0.0002)	(0.0003)
SAT (Reading)	-	-	0.001‡	0.001‡	0.001‡	-	-	0.001‡	0.001‡	0.001‡
			(0.0002)	(0.0002)	(0.0002)			(0.0002)	(0.0002)	(0.0002)
Class Fixed Effects	No	No	No	Yes	Yes	No	No	No	Yes	Yes
Session Fixed Effects	No	No	No	No	Yes	No	No	No	No	Yes
R^2	0.0502	0.1254	0.3829	0.4073	0.4337	0.0748	0.1651	0.4481	0.4714	0.4845

Notes: The dependent variable is the percent of answered questions that were correct ($\overline{y} = 0.784$).
*Statistically significant at the 10-percent level
†Statistically significant at the 5-percent level
‡Statistically significant at the 1-percent level

Table 5 - Results by Sample Traits

Sample Description	Estimated Coefficient on Treatment x Athlete	Sample Size
Full Sample	-0.110†	75
	(0.044)	
Male	-0.151†	34
	(0.067)	
Female	-0.016	41
	(0.061)	
SAT (Math) above median	-0.004	38
	(0.057)	
SAT (Math) below or equal to median	-0.176*	37
	(0.088)	
White Non-Hispanic	-0.109†	51
	(0.045)	
Non-White	-0.080	24
	(0.187)	
Class of 2008 and 2009	-0.088	33
	(0.052)	
Class of 2010 and 2011	-0.068	42
	(0.072)	

Notes: These specifications condition on the available observables as well as class and session fixed effects (i.e., specification [10] in Table 4).

*Statistically significant at the 10-percent level

†Statistically significant at the 5-percent level

‡Statistically significant at the 1-percent level

Table 6 - Results by Outcome Variables

Dependent Variable	Estimated Coefficient on Treatment x Athlete	Dependent Mean (Standard Deviation)
Test Performance	-0.110†	0.786
	(0.044)	(0.11)
Questions Answered	1.84	35.0
	(2.18)	(4.7)
Correct Answers	-2.30	27.5
	(1.95)	(5.2)
Academic Self-Regard	0.26	0.07
	(0.45)	(0.99)
Sports Words	-0.48	1.9
	(0.67)	(1.4)
Self-Doubt Words	0.21	1.9
	(0.67)	(1.4)

Notes: These specifications (n=75) condition on the available observables as well as class and session fixed effects (i.e., specification [10] in Table 4).

*Statistically significant at the 10-percent level

†Statistically significant at the 5-percent level

‡Statistically significant at the 1-percent level